CHAPITRE 8

La régression linéaire simple

Résumé : Les chapitres précédents ont présenté la notion d'intervalle de confiance et de test, et en ont donné divers exemples.

Objectif : Nous étudions ici un modèle statistique d'usage fréquent, voire incontournable, à cause de son efficacité : la régression linéaire. Dans ce chapitre, nous nous contentons d'expliquer une variable quantitative comme fonction affine d'une seule autre variable quantitative.

1. Présentation du modèle

On considère ici des couples de variables. Dans le cas d'un appartement à vendre, il s'agit de sa surface x_j et de son prix y_j.

EXEMPLE 8.1 (Prix d'un appartement). A environnement (quartier ou ville) donné, la surface d'un appartement détermine assez largement son prix ; pas complètement cependant, à cause de la multitude d'autres facteurs à prendre en compte (étage et présence d'un ascenseur, orientation, parking, gardien, année de construction, etc.). On considère la coupure de journal donnée par la figure 1 (datant du début des années 2000, les prix font rêver !). On dispose donc ici de 28 couples (x_j, y_j). On les représente graphiquement à la figure 2. Le but de ce chapitre est, entre autres, de comprendre les résultats numériques de la régression précisés en-dessous du nuage de points. Vous notez, sans doute avec plaisir, qu'ils mettent en jeu des intervalles de confiance et des tests, vos nouveaux amis.

1.1. Modélisation stochastique. Comme toujours, on modélise le problème en supposant que les appartements auxquels nous avons affaire sont un échantillon représentatif de l'ensemble des appartements à vendre sur Paris. Ainsi, on part, pour l'analyse mathématique, du 28-échantillon $(X_1, Y_1), \ldots, (X_{28}, Y_{28})$ i.i.d. selon une certaine loi sur $\mathbb{R}^2$. La première marginale de cette loi indique la répartition des surfaces des appartements du parc immobilier privé, la seconde, celle de leurs prix. Cette loi n'est évidemment pas une loi-produit, puisque la surface a une influence sur le prix.

On veut quantifier et préciser cette influence. On peut écrire

$$Y_j = f_0(X_j) + \varepsilon_j \ ,$$

cela dit que le prix Y_j est la somme de deux facteurs, un facteur dit modélisé ou expliqué $f_0(X_j)$, parce qu'il ne dépend que de la surface X_j, et un autre facteur ε_j dit stochastique ou résiduel, qui englobe tous les autres paramètres. A cause du fameux théorème de la limite centrale flou, on pourra supposer, le moment venu, que les ε_j suivent une loi normale.

Dans ce qui suit, on s'intéressera uniquement aux relations f_0 affines, du type, pour α_0 et β_0 deux réels (les mêmes pour tout l'échantillon),

$$Y_j = \alpha_0 + \beta_0 X_j + \varepsilon_j \ .$$

1. CENSIER, bas de R. Mouffetard, pied-à-terre, 28 m^2, tt confort. Visite vendredi, samedi, dim. 130 000 € à discuter. Facilités	2. CONTRESCARPE, imm. Ancien, pierre de taille, beau duplex caractère, 50 m^2, poutres, refait neuf, 280 000 €
3. R. St-Simon, en pleine verdure, calme, plein soleil, Superbe appt 4p., 106 m^2, cuis. aménagée, s. de bains moderne, chff. cent. Parfait état. Px 650 000 à discuter. Agence s'abstenir. Direct. Propriétaire.	4. RAPP 7P., 196 m^2 standing, 9 fenêtres plein soleil, 800 000 €
5. R. St André-des-Arts, beau liv + chbre, imm. XVIIIe siècle, 55 m^2, 268 000 €	6. 5^e PRES QUAIS, 7 pces, 190 m^2 caractère, standing, 790 000 €
7. GOBELINS, Beau 5p., 110m^2, gd cft, soleil, 500 000 €	8. GOBELINS, et. élevé, calme, asc., 2 pièces, 60 m^2, 320 000 €
9. CENSIER, très grand studio + entrée 48 m^2, tt cft, ensoleillé, calme, bel imm., 250 000 €	10. PANTHÉON, 7^e étage, ascenseur, grand studio 35 m^2 + terrasse. Vue. 250 000 €
11. RUE MADAME, 3P. + Serv., 86 m^2, 350 000 €	12. RUE DE SEINE, 3P., tt cft, 65 m^2, calme, soleil, 300 000 €
13. PANTHEON, bel imm., verdure, magnifique studio 32 m^2, caractère, 155 000 €	14. SEVRES BAB, 1er ét., 2P., gde cuis., bns, 52 m^2, état neuf, 245 000 €
15. MONTPARNASSE, Part. vend atelier d'artiste 40 m^2, duplex, vue imprenable, tout confort, Prix 200 000 €	16. RUE D'ASSAS, imm. gd standing, bel appart 260 m^2, triple récept. + 5 ch., tt cft (travaux) 2 park., 2 ch. Serv., Prix 1 500 000 € à déb.
17. BD St-GERMAIN, 4P., 70 m^2, à amén., 4^e ét., 325 000 €	18. ÎLE St-LOUIS, Lux. appt., 117 m^2, en duplex, gde récept., gde chambre, 2 sdb, Terras., parf. et., décor tr. bon goût, 950 000 €
19. JUSSIEU, Charme, gd 3 pces, 90 m^2, 378 000 €	20. QUARTIER LATIN, 30m^2 à aménager, prix 78 000 €
21. MONTPARNASSE, Imm. p.d.t., 4-5 P., 105 m^2, bon état, 375 000 €	22. RUE MAZARINE, 4^e ét., sans ascens., 52 m^2 à rénover. Prix total 200 000 €
23. CENSIER, Bel imm., 4P. 80 m^2, tt cft, petits travaux, 270 000 €	24. ASSAS LUXEMBOURG, 3P. 60 m^2 s/arbres, imm. caractère, 295 000 €
25. SUR JARDINS OBSERVATOIRE, 140 m^2, grand charme, 990 000 €	26. RUE DE SAVOIE, 4^e ét., Studio 20 m^2, dche, 85 000 €. crédit possible
27. PRES LUXEMBOURG, Bel imm., pierre de taille, Appartement 100 m^2, salon, sal. à manger, 2 chbres, office, cuis., bains, chf. cent., asc., prix : 495 000 €	28. Mo GOBELINS, studio, cuis., s. de bains, 28m^2, calme. Prix 85.000 €

FIG. 1. Liste de 28 appartements à vendre.

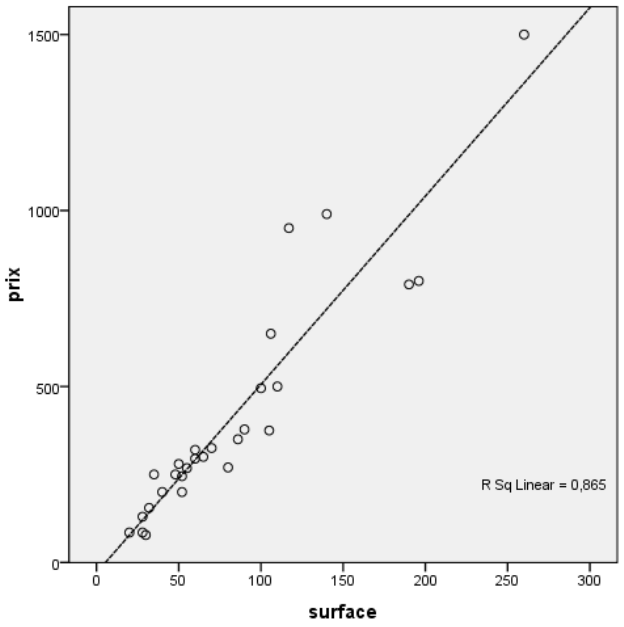
REMARQUE 8.1 (Autres types de dépendance). Quitte à considérer les $\ln X_j$ ou les X_j^2 en lieu et place des X_j, on peut évidemment aussi s'intéresser à des relations comme

$$Y_j = \alpha_0 + \beta_0 X_j^2 + \varepsilon_j \qquad \text{ou} \qquad Y_j = \alpha_0 + \beta_0 \ln X_j + \varepsilon_j \ .$$

REMARQUE 8.2 (Caractère aléatoire ou non du plan d'expérience). On appelle la suite des X_j le plan d'expérience. Il peut être aléatoire comme dans le cas des appartements (on étudie ce qu'on lit dans le journal), ou fixé par l'expérimentateur.

Un autre exemple de plan aléatoire serait la détermination du budget vacances Y_j en fonction du revenu mensuel du foyer X_j. Si l'on appelle des Français au hasard dans l'annuaire, les X_j sont aléatoires.

Un exemple de plan fixé par l'expérimentateur serait le cas de l'étude du rendement Y_j d'un champ en fonction de la quantité d'engrais x_j épandue. On considérerait n champs côte à côte soumis aux mêmes conditions climatiques et on considérerait que les variations



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,930[a]	,865	,860	122,939

a. Predictors: (Constant), surface

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-29,466	41,246		-,714	,481	-114,247	55,316
	surface	5,353	,414	,930	12,931	,000	4,502	6,204

a. Dependent Variable: prix

FIG. 2. Résultat de la régression du prix des appartements par leur surface.

sont causées par les quantités d'engrais. Ici, on se fixerait ces quantités x_j à l'avance, elles ne seraient donc pas aléatoires.

EXEMPLE 8.2 (Pourquoi parle-t-on de régression ?). Sir Galton étudiait la taille des fils y_j en fonction de la taille des pères x_j. Il a noté un retour vers un comportement moyen : les pères grands donnaient naissance à des fils plus petits, et les pères petits donnaient naissance à des fils plus grands. "Regression" signifie en anglais « retour » (vers la moyenne, ici). D'où, vous commencez à vous y habituer, la mauvaise traduction franglaise « régression », désormais synonyme de relation en statistique.

Mathématiquement, on s'attend ici à une relation du type $y_j = m_0 + \beta_0(y_j - m_0) + \ldots$, avec m_0 la taille moyenne de la population et $\ldots$ désignant la réalisation d'une variation aléatoire. C'est bien une relation que l'on peut modéliser sous la forme

$$Y_j = \alpha_0 + \beta_0 X_j + \varepsilon_j \ .$$

Ici, on voudrait estimer β_0 et tester que $0 < \beta_0 < 1$.

2. Estimation dans le cas général (non nécessairement gaussien)

Comme indiqué à la remarque 8.2, et quitte à conditionner par les valeurs observées x_j des X_j, on suppose désormais qu'on est dans le modèle

$$Y_j = \alpha_0 + \beta_0 x_j + \varepsilon_j \ , \qquad j = 1, \dots, n,$$

où $\varepsilon_1, \dots, \varepsilon_n$ sont des variables aléatoires i.i.d. selon une certaine loi centrée, admettant un moment d'ordre deux :

$$\mathbb{E}[\varepsilon_j] = 0 \qquad \text{et} \qquad \mathbb{E}\left[\varepsilon_j^2\right] = \sigma_0^2 \ .$$

On rappelle les seules réalisations que l'on observera sont celles des Y_j, notées y_j. On n'observera pas les réalisations des ε_j (et c'est pour cela que je les appelle variables aléatoires, et non pas observations).

Les paramètres α_0 et β_0 sont inconnus et voudrait les estimer.

2.1. Estimateurs des moindres carrés des coefficients de régression.

Définition 8.1. *On appelle estimateurs des moindres carrés de* (α_0, β_0) *le couple* $(\widehat{\alpha}_n, \widehat{\beta}_n)$ *tel que*

$$(\widehat{\alpha}_n, \widehat{\beta}_n) = \underset{(\alpha,\beta)\in\mathbb{R}^2}{\operatorname{argmin}} \sum_{j=1}^{n} \big(Y_j - (\alpha + \beta x_j)\big)^2 \ .$$

Si l'on note

$$F_n : (\alpha, \beta) \in \mathbb{R}^2 \mapsto \sum_{j=1}^{n} \big(Y_j - (\alpha + \beta x_j)\big)^2 \ ,$$

alors le couple $(\widehat{\alpha}_n, \widehat{\beta}_n)$ est un point critique de F_n, soit

$$\frac{\partial F_n}{\partial \alpha}(\alpha, \beta) = \frac{\partial F_n}{\partial \beta}(\alpha, \beta) = 0 \ .$$

Si l'on faisait les calculs, cela nous donnerait un système de deux équations à deux inconnues (α et β). En le résolvant, on obtiendrait les formules suivantes.

Proposition 8.1. *On définit les quantités empiriques suivantes,*

$$\mathrm{Var}(x_1^n) = \frac{1}{n}\sum_{j=1}^{n} (x_j - \overline{x}_n)^2 \qquad \textit{et} \qquad \mathrm{Cov}(x_1^n, Y_1^n) = \frac{1}{n}\sum_{j=1}^{n} (Y_j - \overline{Y}_n)(x_j - \overline{x}_n) \ ,$$

appelées respectivement variance et covariance empiriques (ou d'échantillon). Alors les estimateurs des moindres carrés de α_0 *et* β_0 *sont donnés par*

$$\widehat{\beta}_n = \frac{\mathrm{Cov}(x_1^n, Y_1^n)}{\mathrm{Var}(x_1^n)} \qquad \textit{et} \qquad \widehat{\alpha}_n = \overline{Y}_n - \widehat{\beta}_n \overline{x}_n \ .$$

La minute SPPS 8.1. Sur la figure 2 (obtenue par `Analyze / Regression / Linear` en ce qui concerne les tableaux de statistiques), on lit les valeurs réalisées pour $\widehat{\alpha}_n$ et $\widehat{\beta}_n$, soit, respectivement, -29.466 et 5.353.

Exercice 8.1. Ceux d'entre vous qui voudraient refaire ces calculs peuvent se reporter au polycopié de cours de l'année 2007, page 156 du fichier `pdf`.

En particulier, la droite de régression est l'ensemble des points

$$\mathcal{D} = \left\{ (x,\, \widehat{\alpha}_n + \widehat{\beta}_n x),\ x \in \mathbb{R} \right\} = \left\{ \left(x,\, \overline{Y}_n + \frac{\mathrm{Cov}(x_1^n,\, Y_1^n)}{\mathrm{Var}(x_1^n)} (x - \overline{x}_n) \right),\quad x \in \mathbb{R} \right\} .$$

On l'a tracée à la figure 2. Elle passe toujours par le point $(\overline{x}_n,\, \overline{Y}_n)$. Elle passe également par les points

$$\widehat{Y}_j = \widehat{\alpha}_n + \widehat{\beta}_n x_j = \overline{Y}_n + \frac{\mathrm{Cov}(x_1^n,\, Y_1^n)}{\mathrm{Var}(x_1^n)} (x_j - \overline{x}_n) \ ,$$

pour $j = 1, \ldots, n$.

Les $\widehat{Y}_j$ sont les valeurs prédites pour les observations Y_j par notre modèle. Notez que lorsque nous formons les $\widehat{Y}_j$, nous connaissons déjà les Y_j : l'expression des $\widehat{Y}_j$ dépend des Y_j. Cependant, à cause de la contrainte de modèle linéaire, l'interpolation ne peut être parfaite et les écarts

$$\widehat{\varepsilon}_j = Y_j - \widehat{Y}_j \ , \qquad j = 1, \ldots, n,$$

seront utilisés pour mesurer l'adéquation du modèle. On les appelle les résidus (ou écarts résiduels). Ils ne sont pas les réalisations des ε_j, mais presque, aux erreurs d'estimations près de α_0 et β_0 par $\widehat{\alpha}_n$ et $\widehat{\beta}_n$.

2.2. Le coefficient de détermination r^2. Remarquez que nous avons pris $\widehat{\alpha}_n$ et $\widehat{\beta}_n$ de telle sorte que la somme des carrés des résidus $\sum (\widehat{\varepsilon}_j)^2$ soit minimale. On note également qu'on a $\sum \widehat{\varepsilon}_j = 0$: certaines des observations Y_j se situent donc au-dessus de la droite de régression tandis que d'autres se situent en-dessous ; et leurs écarts à la droite se compensent. On a également

$$\frac{1}{n} \sum_{j=1} n \widehat{Y}_j = \overline{Y}_n = \frac{1}{n} \sum_{j=1} n Y_j \ .$$

Une conséquence très importante de ces faits est la décomposition suivante.

Théorème 8.1. *La somme des carrés totale Σ_T est égale à la somme des carrés expliquée par la régression Σ_E plus la somme des carrés résiduelle Σ_R,*

$$\underbrace{\sum_{j=1}^{n} (Y_j - \overline{Y}_n)^2}_{\overset{\text{not.}}{=} \Sigma_T} = \underbrace{\sum_{j=1}^{n} (\widehat{Y}_j - \overline{Y}_n)^2}_{\overset{\text{not.}}{=} \Sigma_E} + \underbrace{\sum_{j=1}^{n} (Y_j - \widehat{Y}_j)^2}_{\overset{\text{not.}}{=} \Sigma_R} \ .$$

Σ_T mesure la variabilité des observations Y_j autour de leur moyenne $\overline{Y}_n$; à un coefficient $1/(n-1)$, elle est égale à l'estimateur de la variance. Le terme Σ_R mesure la variabilité des prédictions du modèle $\widehat{Y}_j$ autour de leur moyenne $\overline{Y}_n$, c'est en un sens la variabilité intrinsèque du modèle. Le dernier terme mesure quant à lui la taille des résidus (leur variabilité autour de leur moyenne nulle).

Définition 8.2. *Le coefficient de détermination r^2 est la fraction de la variabilité totale expliquée par la régression,*

$$r^2 = \frac{\Sigma_E}{\Sigma_T} \ .$$

On a donc $0 \leqslant r^2 \leqslant 1$. Le cas limite $r^2 = 1$ exprime une adéquation linéaire parfaite : les écarts résiduels $\widehat{\varepsilon}_j$ sont tous nuls, le modèle linéaire semble parfaitement déterministe, au vu des données recueillies. Au contraire, une faible valeur de r^2 indique une faible liaison linéaire entre les x_j et les y_j ; cela ne signifie pas qu'il n'existe pas de liaison forte entre x_j et y_j : lorsque celle-ci existe, et cela est possible, elle n'est simplement pas linéaire (les y_j peuvent par exemple être linéaires en les x_j^2).

La minute SPPS 8.2. Sur la figure 2, on lit une valeur réalisée de 86.5 % pour r^2. On traduira à l'homme de la rue : la surface explique 86.5 % du prix de l'appartement.

2.3. Quelques qualités des estimateurs $\widehat{\alpha}_n$ et $\widehat{\beta}_n$.

Proposition 8.2. *Les estimateurs $\widehat{\alpha}_n$ et $\widehat{\beta}_n$ sont sans biais,*

$$\mathbb{E}[\widehat{\alpha}_n] = \alpha_0 \qquad \textit{et} \qquad \mathbb{E}\left[\widehat{\beta}_n\right] = \beta_0 \ .$$

Exercice 8.2. Prouvez ce fait. Il suffit par exemple de voir, pour $\widehat{\beta}_n$, que

$$\mathbb{E}\Big[\mathrm{Cov}(x_1^n,\, Y_1^n)\Big] = \frac{1}{n}\sum_{j=1}^{n}(x_j - \overline{x}_n)\,\mathbb{E}\left[Y_j - \overline{Y}_n\right]$$

$$= \frac{1}{n}\sum_{j=1}^{n}(x_j - \overline{x}_n)\left((\alpha + \beta x_j) - (\alpha + \beta\overline{x}_n)\right) = \beta\ \mathrm{Var}(x_1^n) \ .$$

Remarque 8.3. Lorsque le plan d'expérience est aléatoire, i.e., que les x_j sont les réalisations de variables aléatoires X_j, alors les estimateurs considérés sont également consistants,

$$\widehat{\alpha}_n \xrightarrow{\ \mathbb{P}\ } \alpha_0 \qquad \text{et} \qquad \widehat{\beta}_n \xrightarrow{\ \mathbb{P}\ } \beta_0 \ .$$

Cela procède des propriétés générales de la méthode des moments ; je vous renvoie les plus courageux d'entre vous pour les détails éventuels au polycopié de cours de l'année 2007, pages 157–158 du fichier `pdf`.

Enfin, dans tous les cas, on peut montrer que parmi tous les estimateurs sans biais de α_0 et β_0 qui sont fonctions linéaires des Y_j, les estimateurs $\widehat{\alpha}_n$ et $\widehat{\beta}_n$ sont ceux de variance minimale.

3. Estimation et tests dans le cas du modèle linéaire gaussien

On se place désormais dans le cas

$$Y_j = \alpha_0 + \beta_0 x_j + \varepsilon_j \ , \qquad j = 1, \ldots, n,$$

où $\varepsilon_1, \ldots, \varepsilon_n$ sont des variables aléatoires i.i.d. selon une loi normale $\mathcal{N}(0,\, \sigma_0^2)$. C'est ce qu'on appelle un modèle linéaire gaussien. On le justifie par les arguments habituels de théorème de la limite centrale floue, quand une foule de petits facteurs indépendants expliquent les variations par rapport à l'ajustement linéaire ; ces variations sont alors normales.

Ici, comme auparavant, on suppose que le modèle est homoscédastique, i.e., la variance des variations est indépendante des x_j. On a trois paramètres, que l'on va chercher à estimer, encadrer et tester : α_0, β_0 et σ_0^2.

3.1. Lois des estimateurs en jeu. A cause de l'hypothèse formulée sur la loi des ε_j, on va ici pouvoir préciser la loi des estimateurs en jeu.

La proposition 8.1 donne les expressions explicites de $\widehat{\alpha}_n$ et $\widehat{\beta}_n$: tous deux sont des combinaisons linéaires des Y_j. Or, ces dernières sont des variables aléatoires indépendantes, chacune de loi normale, puisque les ε_j le sont également. En conséquence, $\widehat{\alpha}_n$ et $\widehat{\beta}_n$ suivent chacun une loi normale, dont il ne reste plus qu'à déterminer les paramètres. La proposition 8.2 montre qu'ils ont pour espérances respectives α_0 et β_0. Un calcul pas forcément aisé ni agréable donne alors le résultat suivant.

THÉORÈME 8.2. *Dans le modèle linéaire gaussien, les estimateurs $\widehat{\alpha}_n$ et $\widehat{\beta}_n$ suivent les lois normales*

$$\widehat{\beta}_n \sim \mathcal{N}\left(\beta_0, \frac{\sigma_0^2}{n \operatorname{Var}(x_1^n)}\right) \qquad \text{et} \qquad \widehat{\alpha}_n \sim \mathcal{N}\left(\alpha_0, \frac{\sigma_0^2}{n}\left(1 + \frac{(\overline{x}_n)^2}{\operatorname{Var}(x_1^n)}\right)\right) .$$

EXERCICE 8.3. Ceux qui veulent s'entraîner au calcul algébrique intensif peuvent essayer de retrouver à la main ce résultat. Le corrigé se trouve, ici encore, dans le polycopié de cours de l'année 2007, pages 159–160 du fichier `pdf`.

Ce qui concerne l'estimation de la variance σ_0^2 des écarts aléatoires au comportement linéaire est à rapprocher des résultats du paragraphe 6.2. On commence par noter que par définition du modèle linéaire gaussien et de la loi du χ^2,

$$\sum_{j=1}^{n} \varepsilon_j^2 \sim \sigma_0^2 \, \chi_n^2 .$$

Ici, le mieux qu'on puisse faire est de considérer la somme des carrés des écarts résiduels, dont on peut prouver qu'elle suit également une loi du χ^2, mais à $n-2$ degrés de liberté (puisque cette fois-ci, on a dû estimer deux paramètres, α_0 et β_0),

$$\Sigma_R = \sum_{j=1}^{n} (\widehat{\varepsilon}_j)^2 \sim \sigma_0^2 \, \chi_{n-2}^2 .$$

On traduit cela par le théorème suivant.

THÉORÈME 8.3. *Dans le modèle linéaire gaussien, on dispose de l'estimateur de la variance σ_0^2*

$$\widehat{\sigma}_n^2 = \frac{1}{n-2} \sum_{j=1}^{n} (\widehat{\varepsilon}_j)^2 \sim \frac{\sigma_0^2}{n-2} \chi_{n-2}^2 .$$

Il est sans biais et consistant.

Une propriété supplémentaire qui serait utile si on faisait les preuves des différents théorèmes est que est $\widehat{\sigma}_n^2$ est indépendant de $\widehat{\alpha}_n$ et $\widehat{\beta}_n$.

3.2. Estimation et test sur β_0 : la variable explicative influe-t-elle sur la variable à expliquer ? En combinant les résultats du paragraphe précédent, on aboutit à l'égalité en loi suivante, de laquelle se déduisent facilement intervalle de confiance sur β_0 et test de β_0 à une valeur de référence :

$$\sqrt{n \operatorname{Var}(x_1^n)} \, \frac{\widehat{\beta}_n - \beta_0}{\widehat{\sigma}_n^2} \sim \mathcal{T}_{n-2} .$$

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2527207,505	1	2527207,505	167,210	,000[a]
	Residual	392963,209	26	15113,970		
	Total	2920170,714	27			

a. Predictors: (Constant), surface

b. Dependent Variable: prix

FIG. 3. Résultat de l'analyse de la variance sur les données de prix d'appartement.

COROLLAIRE 8.1. *Un intervalle de confiance exact de niveau* $1-p$ *sur* β_0 *est donné par exemple par*

$$\widehat{I}_n = \left[\widehat{\beta}_n \pm t_{n-2,1-p/2}\sqrt{\frac{\widehat{\sigma}_n^2}{n\ \mathrm{Var}(x_1^n)}}\right] \ .$$

PRINCIPE 8.1. *On part d'une situation de modèle linéaire gaussien* $Y_j = \alpha_0 + \beta_0 x_j + \varepsilon_j$*, où les* ε_j *sont i.i.d. selon une loi normale* $\mathcal{N}(0, \sigma_0^2)$ *et on se demande si* β_{ref} *est une valeur admissible pour* β_0*. Le test est fondé sur le résultat suivant : sous* $H_0 : \beta_0 = \beta_{\mathrm{ref}}$*,*

$$T_n = \sqrt{n\ \mathrm{Var}(x_1^n)}\ \frac{\widehat{\beta}_n - \beta_{\mathrm{ref}}}{\widehat{\sigma}_n^2} \ \sim\ \mathcal{T}_{n-2} \ .$$

en particulier, la distribution de T_n *est centrée autour de 0. Lorsque* β_0 *est plus grand que* β_{ref}*,* T_n *tend à prendre des valeurs plus grandes (que 0). Lorsque* β_0 *est plus petit que* β_{ref}*,* T_n *tend à prendre des valeurs plus petites (que 0).*

L'usage le plus fréquent de ce test est avec l'hypothèse $H_0 : \beta_0 = 0$ d'absence de liaison linéaire entre les x_j et les Y_j. Dans ce cas, un calcul simple montre que la statistique de test T_n introduite dans le principe précédent est telle que

$$F = (T_n)^2 = \left(\sqrt{n\ \mathrm{Var}(x_1^n)}\ \frac{\widehat{\beta}_n}{\widehat{\sigma}_n^2}\right)^2 = (n-2)\,\frac{\sum_{j=1}^n (\widehat{Y}_j - \overline{Y}_n)^2}{\sum_{j=1}^n (Y_j - \widehat{Y}_j)^2} = (n-2)\,\frac{r^2}{1-r^2} \ .$$

La statistique de test T_n est donc une fonction de r^2 ; que r^2 joue un rôle ne doit pas vous surprendre au vu des commentaires qui suivent la définition 8.2.

On généralisera dans le cas de la régression multiple ces tests d'existence ou d'absence de relation linéaire fondés sur r^2. On retiendra qu'ici on a comparé la meilleure approximation $\overline{Y}_n$ des moindres carrés des Y_j sous l'hypothèse $H_0 : \beta_0 = 0$ à celles proposés sous H_1, et qui sont les $\widehat{Y}_j$.

3.3. Estimations et tests sur α_0 et σ_0^2. Ils procèdent des égalités en distribution

$$\widehat{\alpha}_n \ \sim\ \mathcal{N}\left(\alpha_0,\ \frac{\sigma_0}{n}\left(1 + \frac{(\overline{x}_n)^2}{\mathrm{Var}(x_1^n)}\right)\right) \qquad \text{et} \qquad \widehat{\sigma}_n^2 = \frac{1}{n-2}\sum_{j=1}^n (\widehat{\varepsilon}_j)^2 \ \sim\ \frac{\sigma_0^2}{n-2}\,\chi^2_{n-2}$$

selon une mécanique qui devrait vous sembler bien familière désormais. Vous pouvez les définir plus précisément en exercice. Nous nous contenterons de lire les valeurs calculées par SPSS.

3.4. Décryptage des sorties SPSS. Dans ce paragraphe, nous allons apprendre à lire les sorties SPSS.

LA MINUTE SPPS 8.3. On commence par le premier tableau de la figure 2. A gauche, on lit respectivement les valeurs de $\pm\sqrt{r^2}$ (avec le signe de $\widehat{\beta}_n$) et de r^2, ici, 0.930 et 0.865. A droite, on lit l'estimée correspondant à la racine carrée de l'estimateur de la variance $\widehat{\sigma}_n^2$, soit 122.939 ici.

LA MINUTE SPPS 8.4. On continue par le second tableau de la figure 2. Dans la colonne la plus à gauche, on lit les estimées de α_0 (haut) et β_0 (bas), qui valent respectivement -29.466 et 5.353. La seconde colonne donne les valeurs des estimées de la variance des estimateurs respectifs, à savoir, les valeurs de

$$\frac{\widehat{\sigma}_n^2}{n}\left(1+\frac{(\overline{x}_n)^2}{\mathrm{Var}(x_1^n)}\right) \qquad \text{et} \qquad \frac{\widehat{\sigma}_n^2}{n}\,\mathrm{Var}(x_1^n)\ ;$$

ce sont elles, on l'a vu, qui donnent la demie-longueur des intervalles de confiance sur α_0 et β_0. Les estimées valent respectivement 41.286 et 0.414. La quatrième colonne donne la valeur des statistiques de test suivant la loi de Student, celles rappelées au début des paragraphes 3.2 et 3.3 ; elles valent -0.714 et 12.931, respectivement. La cinquième colonne donne la P-valeur associée à un test de significativité bilatère ($H_0 : \alpha_0 = 0$ vs. $H_1 : \alpha_0 \neq 0$, et de même pour β_0). Ici, on lit respectivement 48.1 % ($\alpha_0 = 0$ est plausible) et 0.000, ce qui signifie que cette seconde P-valeur est, vu les arrondis, plus petite que 5×10^{-4} (l'hypothèse $H_0 : \beta_0 = 0$ est en revanche clairement rejetée). Enfin, les deux dernières colonnes donnent les intervalles de confiance sur α_0 et β_0, au niveau fixé par l'utilisateur lorsqu'il lance la régression. Notez que le premier contient 0 mais pas le second.

LA MINUTE SPPS 8.5. On passe enfin à la figure 3. Dans la première colonne on lit, de haut en bas, σ_E, σ_R et Σ_T. La deuxième colonne ne nous intéressera que dans le chapitre suivant, sur la régression multiple. Les éléments de la troisième colonne sont ceux de la première divisés par ceux de la seconde ; nous verrons au chapitre suivant, là aussi, pourquoi cela est intéressant. La statistique F de la quatrième colonne a été définie plus haut (comme $(T_n)^2$) ; elle suit en fait une loi de Fisher (ici, à 1 et 26 degrés de liberté). On peut tester $H_0 : \beta_0 = 0$ avec F et on obtient un test équivalent à celui avec la loi de Student ; et sans surprise, on lit donc la même P-valeur.

EXERCICE 8.4. On a vu que l'on pouvait prendre $\alpha_0 = 0$. Lancez une régression avec coefficient constant nul (par `Regression / Linear / Options` puis en décochant `Include constant in equation`). Montrez que l'on propose alors la relation "prix $= 5.109 \times$ surface + aléa".

4. Estimation et prédiction en un nouveau point (dans le cas du modèle linéaire gaussien)

On considère toujours le modèle linéaire gaussien

$$Y_j = \alpha_0 + \beta_0 x_j + \varepsilon_j\ , \qquad j = 1, \ldots, n,$$

où $\varepsilon_1, \ldots, \varepsilon_n$ sont des variables aléatoires i.i.d. selon une loi normale $\mathcal{N}(0, \sigma_0^2)$. On a indiqué aux paragraphes précédents comment ajuster le modèle aux données (comment estimer α_0, β_0, et σ_0^2, et comment déterminer s'il existe ou pas une relation linéaire).

Soit un nouveau point x, déterminé par le dispositif expérimental ou choisi par l'utilisateur. On note $Y_x = \alpha_0 + \beta_0 x + \varepsilon_x$ l'observation associée (où $\varepsilon_x \sim \mathcal{N}(0,\, \sigma_0^2)$ est indépendante des variables aléatoires précédentes) ; son espérance est notée $\mu_x = \alpha_0 + \beta_0 x$. On peut vouloir, dans le cadre de la prévision,

– donner un intervalle de confiance sur μ_x, l'espérance de Y_x,
– ou même, donner un intervalle dans lequel Y_x sera avec grande probabilité.

4.1. Intervalle de confiance sur l'espérance μ_x. Certes, les techniques décrites au paragraphe 3 indiquaient des intervalles de confiance sur α_0 et β_0, et s'ils sont simultanément vrais, alors on peut en déduire un intervalle de confiance sur $\mu_x = \alpha_0 + \beta_0 x$. L'intervalle qu'on obtiendrait ainsi (exercice : explicitez-le !) a pour défaut que sa demie-longueur croît rapidement avec x. On voudrait une formule moins sensible à la valeur de x.

On va procéder de la même manière qu'au paragraphe 3.1. Les estimateurs $\widehat{\alpha}_n$ et $\widehat{\beta}_n$ sont combinaisons linéaires des observations gaussiennes indépendantes Y_j ; l'estimateur

$$\widehat{\mu}_{x,n} = \widehat{\alpha}_n + \widehat{\beta}_n\, x$$

est lui aussi une telle combinaison linéaire et suit par conséquent une loi normale (dont il suffit de déterminer espérance et variance). L'espérance est bien μ_x, vu le caractère sans biais de $\widehat{\alpha}_n$ et $\widehat{\beta}_n$. Des calculs similaires à ceux proposés à l'exercice 8.3 permettent de déterminer la variance, puis, moyennant une studentisation, conduisent au résultat suivant.

THÉORÈME 8.4. *Dans le modèle linéaire gaussien, l'estimateur* $\widehat{\mu}_{x,n} = \widehat{\alpha}_n + \widehat{\beta}_n\, x$ *suit une loi normale,*

$$\widehat{\mu}_{x,n} \;\sim\; \mathcal{N}\left(\mu_x,\; \frac{\sigma_0^2}{n}\, h_{x,n}\right) \qquad \text{où} \quad h_{x,n} = 1 + \frac{1}{\mathrm{Var}(x_1^n)}\,(x - \overline{x}_n)^2$$

est le levier en x.

Par conséquent,

$$\sqrt{\frac{n}{\widehat{\sigma}_n^2\, h_{x,n}}}\;\widehat{\mu}_{x,n} \;\sim\; \mathcal{T}_{n-2}\ .$$

On déduit de la dernière assertion des intervalles de confiance et des tests sur μ_x, selon la cuisine habituelle. Par exemple, un intervalle de confiance exact de niveau $1-\alpha$ sur μ_x est

$$\left[\widehat{\alpha}_n + \widehat{\beta}_n\, x \;\pm\; t_{n-2,1-\alpha/2}\sqrt{\frac{\widehat{\sigma}_n^2}{n}\, h_{x,n}}\right]\ .$$

EXERCICE 8.5. Ceux qui veulent s'entraîner au calcul algébrique intensif peuvent ici encore essayer de calculer la variance de $\widehat{\mu}_{x,n}$. Le corrigé se trouve là aussi dans le polycopié de cours de l'année 2007, pages 168–169 du fichier `pdf`.

4.2. Intervalle de prévision sur Y_x. On cherche maintenant à donner des indications sur la valeur non plus de μ_x, mais de $Y_x = \mu_x + \varepsilon_x$. Il suffit de tenir compte de l'ajout de la perturbation aléatoire ε_x, qui est indépendante des observations $Y_1, \ldots, Y_n$ (et des aléas correspondants, $\varepsilon_1, \ldots, \varepsilon_n$).

On cherche un intervalle $\widehat{I}_{x,n}$ tel que

$$\mathbb{P}\{Y_x \in \widehat{I}_{x,n}\} \geqslant 1 - \alpha$$

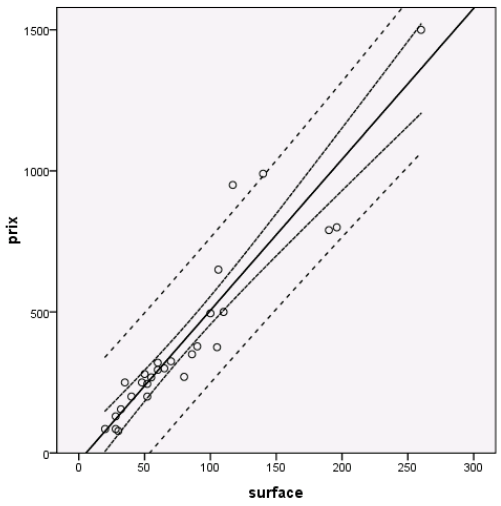


FIG. 4. Intervalles de confiance et de prédiction sur l'exemple des appartements.

pour un niveau $1-\alpha$ fixé par l'utilisateur. $\widehat{I}_{x,n}$ est appelé un intervalle de prévision, plutôt qu'un intervalle de confiance. En effet, les intervalles de confiance portent sur des quantités déterministes ; quand on veut encadrer une quantité aléatoire qui sera observée plus tard, on parle, comme ici, d'intervalle de prévision.

On a de bons espoirs d'aboutir : on a un intervalle de confiance sur μ_x et on a une estimation $\widehat{\sigma}^2_n$ de la variance σ^2_0 de $\varepsilon_x \sim \mathcal{N}(0,\, \sigma^2_0)$. Des calculs simples, utilisant le résultat du théorème 8.4, montrent alors le résultat suivant.

THÉORÈME 8.5. *Dans le modèle linéaire gaussien, la différence entre l'observation* Y_x *et la prédiction de son espérance* $\widehat{\mu}_{x,n}$ *suit une loi normale,*

$$Y_x - \widehat{\mu}_{x,n} \;\sim\; \mathcal{N}\left(0,\; \sigma^2_0\left(1+\frac{h_{x,n}}{n}\right)\right) \;.$$

Une studentisation donne alors

$$\frac{1}{\sqrt{\widehat{\sigma}^2_n\,(1+h_{x,n}/n)}}\,(Y_x-\widehat{\mu}_{x,n}) \;\sim\; \mathcal{T}_{n-2}$$

et par conséquent, un intervalle de prévision au niveau $1-\alpha$ *est*

$$\widehat{I}_{x,n} = \left[\widehat{\mu}_{x,n} \;\pm\; t_{n-2,1-\alpha/2}\sqrt{\widehat{\sigma}^2_n\left(1+\frac{h_{x,n}}{n}\right)}\right] \;.$$

EXERCICE 8.6. Cette fois-ci, vous avez tout pour prouver la première assertion du théorème ! Le calcul est simple, si l'on exploite le théorème 8.4 et les propriétés d'indépendance des ε_j et de ε_x.

4.3. Tracés SPSS.

LA MINUTE SPPS 8.6. En traçant un `scatterplot` (par `Chart Builder / Scatter/Dot`), et en demandant dans la fenêtre de `Chart editor` le tracé de la droite de régression, des intervalles de confiance et de prévision (tous deux à 95 %), on a obtenu la figure 4.

On remarque qu'il semble y avoir des observations atypiques (qui n'appartiennent pas à leur propre intervalle de prévision). Il faudrait les enlever du jeu de données et expliquer pourquoi (par des considérations extra-statistiques) elles méritent un traitement à part.

5. Exercices

Les exercices qui suivent nécessitent tous des fichiers de données, disponibles, comme à l'accoutumée, sur le site web du cours.

Exercice 8.7. Retrouvez les résultats du rapport de régression du dernier exercice de l'examen de rattrapage 2007, en étudiant sous SPSS le fichier `Orange.sav`. La première colonne est l'étiquette de l'oranger considéré, la deuxième donne l'âge de l'arbre en jours, et la troisième, sa circonférence mesurée (en millimètres). Répondez ensuite aux questions posées lors de l'examen.

Exercice 8.8 (Regarder la télé rend-il fou ?). Chargez le fichier de données `TV.sav` sous SPSS. Il reporte chaque année (première colonne) le nombre de téléviseurs en service (en milliers, deuxième colonne) et le taux de malades mentaux (nombre pour mille habitants, troisième colonne). L'étude a été réalisée en Grande-Bretagne.

- Quelle est d'après vous la variable à expliquer et la variable explicative ?
- Effectuez la régression sous SPSS. Quelle relation proposez-vous ?
- Quelle quantité mesure la bonne explication de la variable à expliquer par la variable explicative ? Que vaut-elle ici ? Est-ce une valeur satisfaisante ?
- Faut-il en conclure que regarder la télé rend fou, et si non, pourquoi ?
- On pourra se convaincre qu'il faut répondre non à la question précédente en étudiant la régression du taux de malades mentaux par l'indice de l'année. Que se passe-t-il donc ?

Exercice 8.9. Le fichier `Ozone.sav` indique, pendant cinq mois, de mai à septembre 1973, les mesures de différentes quantités : la concentration d'ozone, l'intensité du rayonnement solaire, le vent, la température. Plus précisément, voici ce qu'en dit la source :

> Daily readings of the following air quality values for May 1, 1973 (a Tuesday) to September 30, 1973.
>
> - Ozone : Mean ozone in parts per billion from 1300 to 1500 hours at Roosevelt Island
> - Solar.R : Solar radiation in Langleys in the frequency band 4000Ũ7700 Angstroms from 0800 to 1200 hours at Central Park
> - Wind : Average wind speed in miles per hour at 0700 and 1000 hours at LaGuardia Airport
> - Temp : Maximum daily temperature in degrees Fahrenheit at La Guardia Airport.
>
> The data were obtained from the New York State Department of Conservation (ozone data) and the National Weather Service (meteorological data).

Montrez que la température seule n'est pas une bonne explication de la concentration en ozone. Est-ce parce qu'elle n'est pas une variable influant significativement sur la concentration d'ozone ? Précisez la taille de l'aléa associé à cette régression (i.e., l'estimée de la variance) ; commentez sa valeur.

Parmi les trois variables explicatives, quelle est la meilleure prise isolément pour expliquer la concentration ?

Exercice 8.10. Méditez la figure 5 ; notez en particulier que la régression (simple ou multiple) sera très utilisée dans vos cours de finance ! Enfin, si la finance existe encore d'ici là.

Table IV

Predictability Regressions

$$\sigma_{t,T} = a + b\hat{\sigma}_t^T + \varepsilon_{t,T}$$

where $\sigma_{t,T}$, the volatility over the remaining life of the option contract, is regressed against the volatility forecast $\hat{\sigma}_t^T$. This includes the implied standard deviation (ISD) from option prices, a moving average (MA) with a moving window of 20 trading days, and the GARCH time-series model, which is the conditional volatility for the next day based on parameters in Table II. Periods end on February 28, 1992, and start in January 1985 (DM), July 1986 (JY), and March 1985 (SF). Regressions use daily observations, and standard errors are corrected for the induced overlap and heteroskedasticity using the Hansen-White (HW) procedure. Asymptotic HW standard errors in parentheses.

		Slopes On			
Currency	a	ISD	MA(20)	GARCH	R^2
DM	0.323*	0.547*†			0.1564
	(0.115)	(0.138)			
	0.602*		0.190†		0.0540
	(0.084)		(0.099)		
	0.366			0.478*†	0.0499
	(0.191)			(0.227)	
	0.303*	0.669*	−0.099		0.1632
	(0.112)	(0.165)	(0.101)		
	0.401*	0.622*		−0.173	0.1599
	(0.152)	(0.167)		(0.201)	
JY	0.327*	0.496*†			0.0965
	(0.118)	(0.181)			
	0.563*		0.134†		0.0223
	(0.074)		(0.102)		
	−0.063			1.017*	0.0495
	(0.323)			(0.458)	
	0.322*	0.578*	−0.073		0.1004
	(0.117)	(0.204)	(0.111)		
	0.042	0.421*		0.474	0.1051
	(0.289)	(0.177)		(0.399)	
SF	0.392*	0.520*†			0.1454
	(0.149)	(0.175)			
	0.658*		0.182†		0.0542
	(0.087)		(0.099)		
	0.250			0.650*	0.0581
	(0.267)			(0.305)	
	0.370*	0.647*	−0.097		0.1521
	(0.146)	(0.187)	(0.090)		
	0.526*	0.609*		−0.240	0.1490
	(0.210)	(0.201)		(0.262)	

* Significantly different from zero at the 5 percent level.
† Significantly different from unity at the 5 percent level.

FIG. 5. Un extrait d'un article de finance (envoyé par Christophe Pérignon).